6N-09

文章生成 AI を活用した OSS コミュニティへの投稿支援システムの開発

西原　啓亮†　　中才　恵太朗†

大阪公立大学工業高等専門学校総合工学システム学科†

1. はじめに

オープンソースソフトウェア（OSS）コミュニティには不具合の報告や要望等が日々投稿されており，その投稿によりソフトウェアを改善することができるため，積極的な投稿が期待される．ただし，投稿内容の確認が不十分であると，重複した内容や不十分な内容となり，解決に要する時間が長くなる．このような不完全な投稿は，OSS コミュニティの負担となる．

本研究では，手軽に使える汎用の大規模言語モデル（LLM）を用いて，OSS コミュニティへの投稿内容から検索用ラベルや専門用語の抽出，投稿内容の要約を行い，それらを既出の投稿検索や投稿内容の情報補完等に活用する．そして，ソフトウェア開発の理解が浅い人であっても十分に役立つ投稿が可能な OSS コミュニティへの投稿に特化した支援システムを提案する．また，実際の OSS コミュニティに投稿されたデータと検索ラベルを利用し，本提案システムの評価を行った．

2. 言語生成 AI の活用や日本語解析の関連研究

近年、LLM を用いた Chat GPT 等のチャット形式で利用できる言語生成 AI が注目を浴びている．汎用的かつ，手軽に扱いやすいため急速に普及し，その活用例の紹介や成果が日々報告されている．言語生成 AI は，汎用的なものから任意のタスクに特化したモデルまで様々なものがあるが，汎用的なモデルの GPT 3.5 を用いて情報のラベル付けを行う研究がある[1]．ここでは GPT3.5 の API を用いて無視できるコストで合理的なクラスターの粒度を提案できることを示されている．一方で，埋め込みモデル自体に出力が依存する点について指摘されている．

また，日本語を読みやすくする技術として，助詞，格助詞を参照して依存ネットワークを用いて要約する手法が提案されている．[2]

Development of a Support System for Posting to OSS Communities Using Text Generation AI

†Keisuke Nishihara, Department of Technological Systems, Osaka Metropolitan University College of Technology

†Keitaro Nakasai, Department of Technological Systems, Osaka Metropolitan University College of Technology

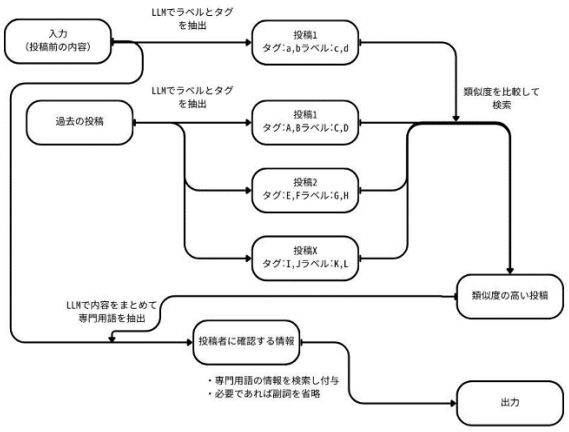


図１　提案システム概要

3. 提案システム概要

提案システムの概要を図 1 に示す．はじめに投稿前の内容と過去の投稿に対して LLM を用いてタグとラベルを抽出し，それぞれ保存する．なお，事前に定義されているものがラベル．検索する際に検索クエリとして活用するものをタグとしている．

次にそれらの類似度を比較することで投稿前の内容と類似度の高い過去の投稿を検索し，再度，投稿前の内容とした検索結果の類似度の高い過去の投稿を LLM で処理し，投稿者に確認する情報として作成する．その情報のうち，専門用語を抽出して説明を付与し，情報をわかりやすくするため，助詞を参照し副詞の省略を行う．

3.1. LLM を用いたタグ，ラベル付けと検索

LLM にプロンプトと投稿の内容を入力しタグとラベルを数値の重み付きで抽出を行う．

LLM にプロンプトを用いて言語処理を行う際は，プロンプトに入力例と出力例のサンプルデータを入れることで質の高い結果が得られる[1]．サンプルデータありのプロンプトを用いて GPT-3.5-turbo モデルに API 経由で処理を行った．

検索は，投稿前の内容の任意の数のTag_1と過去の投稿に付けられたTag_xを掛け合わせたものの和でマッチ度M_xをそれぞれ導出し，これを用いて検索を行う．$M_x = \sum Tag_1 \times Tag_x$

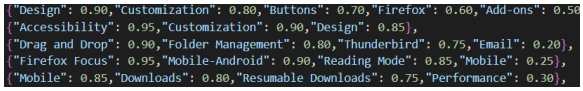
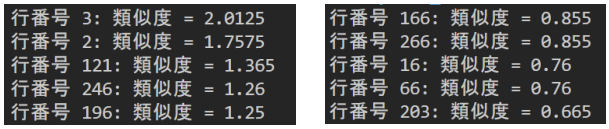


図 2　ラベル付けの実行例



(a)タグによる検索　　(b) ラベルによる検索

図 3　類似度の検索

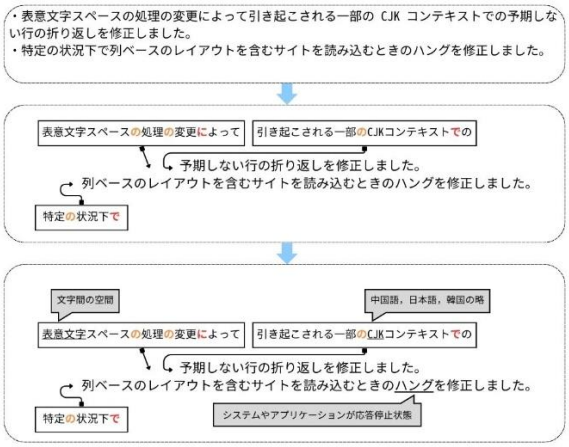


図 4　読みやすいテキストの提示例

3.2.　読みやすいテキストの提示

専門用語を前項と同様に LLM で抽出しその説明を付与することで理解しやすいテキストを作成することが出来る．

一方で，文章構造による理解しやすさという点で考えると，リリースノートなど詳細な情報のある文章では，日本語の特性により長い副詞が前に来て読みにくくなる問題がある．これを解決する方法として日本語の文章を助詞と句読点で区切り，格助詞を解析して副詞を検出し読みやすいテキストを提示する手法を提案する．参照すべき，格助詞「二格」「デ格」「ノ格」の用法は以下の通りである．

「二格」：人やモノの存在場所，所有者，位相の着点，動作の相手・対象，状態の対象，原因，移動動作の目的，状態のときを表す用法

「デ格」：出来事・動作の場所，道具・手段，材料，原因，範囲，限度，基準，動作の主体を表す用法

「ノ格」：所有・所属関係，名詞化．同格の関係，説明，限定，原因，理由，帰属を表す用法

以上の説明から「二格」と「デ格」は副詞にかかることが多いといえる．また，「ノ格」はこれらの副詞に帰属することが多いため，「ノ格」の帰属関係も確認する．

4.　評価

LLM を用いたタグとラベル付け，及びそれらを用いた検索の評価を行うために，OSS コミュニティの Mozilla Connect に投稿された直近 300 個の投稿データと準備した類似内容の投稿 3 つを混ぜて処理を行い，類似した内容を検索によって検索可能か実験を行った．タグは主に検索に使われ，本来ラベルはソートや統計等に使われるものであるが，ラベルでも検索は可能か検証した．

検索結果のマッチ度の高い上位 5 個を図 3 に示す．1, 2, 3 行は類似の内容であり上位 5 個を表示した．タグによる検索では 2 行目と 3 行目を検索によりヒットしているが，ラベルではヒットできず有用性は示されなかった．また，タグによる検索結果では全く異なる内容の行番号 121, 246, 196 がそれぞれ高いマッチ度を示しているがこれらは，LLM がタグの数値を高く割り当てたため高いマッチ度を出したことが分かっている．また，303 個の投稿のうち 8 個は期待したフォーマットにならずエラーとなっている．従って LLM でタグとラベル付けを行う際，それが妥当であるか判断するシステムが必要であるといえる．

また，読みやすいテキストの提示の評価は実際の OSS のリリースノートに提案手法の処理を行い評価した．形態素解析は MeCab を用いて行い助詞を参照することで副詞を分ける．その後，LLM は同様に GPT-3.5-Turbo を使用し，専門用語を抽出し解説を行った．処理より得られた結果を図 4 に示す．実験を行ったテキストでは長い副詞を省略し要点を明確に示し，語句の選定も解説も適切な結果が得られた．

5.　おわりに

用意した実際のデータに対して提案システムの有益性を示すことが出来た．

一方でタグ付けを行う際，それが妥当であるか判断するシステムの作成，及びデータを増やしたうえで，使いやすいシステムであるか，評価を行うことを今後の展望とした．

参考文献

[1] Yuwei Zhang, Zihan Wang, Jingbo Shang.: ClusterLLM: Large Language Models as a Guide for Text Clustering, Proc. 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023), pp.13903–13920 (2023).

[2] 山脇拓, 中野滋徳, 足立顕, 牧野武則：依存ネットワークをもとにしたパラグラフ要約, 情報処理学会研究報告, 自然言語処理(NL), 2003-NL-160, No. 23, pp.37-42(2004).